OPTIMAL FRACTIONAL FACTORIAL SPLIT-PLOT DESIGNS FOR MODEL SELECTION

by

Joslin Goh
B.Math., University of Waterloo, 2007

a Project submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the Department
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
Fall 2009

APPROVAL

Name: Joslin Goh

Degree: Master of Science

Title of Project: Optimal Fractional Factorial Split-Plot Designs for Model Selection

Examining Committee: Dr. Richard Lockhart
Chair

Dr. Derek Bingham, Senior Supervisor

Dr. Boxin Tang, Supervisor

Dr. Jason Loeppky, External Examiner
University of British Columbia, Okanagan

Date Approved: _____

Abstract

Fractional factorial designs are used widely in screening experiments to identify significant effects. It is not always possible to perform the trials in a complete random order and hence, fractional factorial split-plot designs arise. In order to identify optimal fractional factorial split-plot designs in this setting, the Hellinger distance criterion (Bingham and Chipman (2007)) is adapted. The approach is Bayesian and directly incorporates common experimenter assumptions. By specifying prior distributions for the model space, the criterion for fractional factorial split-plot designs aims to discriminate between the most probable competing models. Techniques for evaluating the criterion and searching for optimal designs are proposed. The criterion is then illustrated through a few examples with further discussion on the choice of hyperparameters and flexibility of the criterion.

Keywords: Bayesian; Hellinger distance; Optimal designs; Randomization restriction; Screening

Dedication

To my grandparents

Acknowledgments

I would like to begin by expressing appreciation to my supervisor, Dr. Derek Bingham, for his continuing patience, guidance and support. Many thanks to my committee members, Dr. Boxin Tang and Dr. Jason Loeppky, for their insightful comments and suggestions that led to significant improvement in this project.

There are many friends whom I have made in the past two years. I cannot thank each of you enough for the friendships. Special thanks to Ryan Lekivetz, Chunfang Lin, Flavia Moser and Matt Pratola who have helped me a lot at the final stage of completing the project.

A big hug to Michelle Lim, a great friend who has supported me all along, despite being a complete stranger to statistics. I am also grateful to Woei Chet Lim who has shared his experience as a researcher over the past few years.

Finally, I would like to thank my parents, who have always made me their pride.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Designed experiments are commonly used in the initial stage of the experimentation to identify potential active factors on a process or product. By varying the factors of interest over specified level settings in random order, experimenters gain insights into important factors. Due to limited resources, cost constraints or time restrictions, only a fraction of all possible combinations of level settings can be performed. In order to identify influential effects among the large number of effects, fractional factorial (FF) designs are often used.

It is not always possible to perform the experimental trials in a completely random manner. Reasons involving costs, nature of the process or experimental venue result in FF designs with randomization restrictions and hence, fractional factorial split-plot (FFSP) designs arise.

Like FF designs, FFSP designs can be broadly classified into two types: regular and non-regular. Regular FFSP designs are most common and widely used. Regular FFSP designs are specified in terms of a defining contrast subgroup. In this case, factorial effects are fully aliased with other effects (for lower order terms, preferably with higher order interaction). Typically, significant effects are identified using half-normal probability plots (Daniel (1959)).

Regular FF designs can be ranked using, for example, the minimum aberration (MA) criterion (Fries and Hunter (1980)) that relies heavily on the following assumptions:

A1. Effect Sparsity: There are relatively few significant effects.

A2. Effect Hierarchy: Lower order effects are more likely to be significant than higher order effects. Effects of the same order are equally important.

The word length pattern is the sequence of the length of words in the defining contrast subgroup for a regular design. Short words in the defining contrast subgroup are not desirable since this results in the aliasing of lower order effects with other lower order effects. The MA criterion sequentially minimizes the elements of the word length pattern of a design.

Identifying the "best" regular FFSP designs takes some care. Factors that cause the restrictions are conventionally known as whole-plot (WP) factors while the remaining factors are known as subplot (SP) factors. WP effects, and those effects aliased with WP effects, will have a larger variance than SP effects. Thus, the ranking of such designs typically takes into account the different effect variances. Catalogues for FFSP designs ranked by the MA criterion can be found in Huang et al. (1998), Bingham and Sitter (1999) and Bingham and Sitter (2001). Mukerjee and Fang (2000) introduced the Minimum Secondary Aberration to distinguish between competing non-isomorphic MA FFSP designs. Kulachi et al. (2006) ranked the designs based on the maximum number of clear two factor interactions (2FI's).

Non-regular designs, on the other hand, do not have an associated defining contrast subgroup. A key feature of non-regular designs is that some of the effects are partially aliased. Several methods have been proposed to analyze designs with complex aliasing that entertain models with both main effects and 2FI's (Hamada and Wu (1992); Chipman et al. (1997); Miller and Sitter (2001)). The approaches taken to analyze data arising from these designs rely heavily on a third assumption:

A3. Effect Heredity: An interaction factor is more likely to be significant if one or more of its parents are also significant.

With the additional assumption of effect heredity, non-regular designs are often able to consider both significant main effects and 2FI's. Although non-regular designs are not as common as regular designs, they have received considerable attention in recent years. In the case of FF designs, Deng and Tang (1999) and Tang and Deng (1999) proposed the G-aberration and G_2-aberration to rank non-regular FF designs. The minimum G-aberration (Deng and Tang (1999)) is a generalization of the MA criterion to rank non-regular designs. Maximum estimation capacity (MEC) (Cheng and Mukerjee (1998) and Cheng et al. (1999)), which incorporates belief during model selection, is used to measure the optimality of designs and rank the designs accordingly. Under the MEC criterion, non-regular FF designs often have a higher MEC than competing regular FF designs. Other criteria used to rank non-regular FF designs includes the minimum moment aberration (Xu (2003)), moment aberration projection (Xu and Deng (2005)) and the W_k^r and $\widetilde{W}_k^r$-criteria (Cheng and Tsai (2009)).

In the case of FFSP designs, however, there are few general criteria available to rank non-regular FFSP designs. Generally, the D-optimality criterion is used to rank designs (Goos and Vandebroek (2003); Kulachi and Bisgaard (2005); Kowalski (2002)).

Motivated by the approach taken by Bingham and Chipman (2007) for ranking FF designs, this work focuses on a new methodology to rank both regular and non-regular FFSP designs for any run sizes. Bingham and Chipman (2007) used the Hellinger distance criterion (HD-criterion) that measures the Hellinger distance between predictive densities of all competing models. Through a Bayesian setup, the criterion discriminates models according to the assumptions A1 – A3. The proposed approach can be implemented for any run size and can incorporate prior knowledge of the experimenters. It is similar in spirit to the Kullback-Liebler criterion suggested by Meyer et al. (1996) but is more computationally efficient.

The aim of this work is to adapt the HD-criterion in order to construct optimal FFSP designs. The new criterion developed can be used for both regular and non-regular FFSP

designs of different run sizes. Next, Chapter 2 describes motivational examples for FFSP design. Chapter 3 outlines the model, priors and HD-criterion and Chapter 4 covers the search algorithm for optimal designs. Chapter 5 illustrates two examples, which show the effectiveness of the criterion, while Chapter 6 contains further discussion on the hyperparameters and flexibility of the new criterion. Chapter 7 concludes with a brief summary of the methodology and discusses possible future work.

Chapter 2

Motivating Examples

In this chapter, we will consider two examples of how FFSP designs arise in practice. In addition, the differences between FFSP and FF designs are highlighted.

2.1 Example 1

Consider the thin-film coating process described in Huang et al. (1998). The application involved coating a set of substrates with chemicals, which required two stages to complete the process. The first stage was a batch operation where four substrates were put into the vessel and were coated with three types of chemical, $A - C$. At the second stage, the substrates were removed from the vessel and coated with four other chemicals, $P - S$. Each factor had two levels to be considered: "High" $(+1)$ and "Low" (-1).

The experimenters were interested in the impact of the factors on the properties of the thin-film coating. Since the batch operation in stage one limited the randomization order required in a two-level FF design, a two-level FFSP design was more suitable. $A - C$ are called the WP factors and $P - S$ are known as the SP factors.

The columns of the design matrix in Figure 2.1(a) indicate the levels of the factors that were set for each run of the experiment. Let (1) denote the mean effect. If we assume that

only four factorial effects A, B, P and AP are significant, then the collection of significant effects forms a model with model matrix in Figure 2.1(b).

A	B	C	P	Q	R	S
-1	-1	-1	-1	-1	-1	-1
			1	1	1	-1
			-1	1	-1	1
			1	-1	1	1
1	-1	1	1	1	1	1
			-1	-1	-1	1
			-1	1	-1	1
			1	-1	1	1
1	1	-1	-1	-1	-1	-1
			1	1	1	-1
			-1	1	-1	1
			1	-1	1	1
-1	1	1	-1	-1	-1	-1
			1	1	1	-1
			-1	1	-1	1
			1	-1	1	1

(a) Design matrix

(1)	A	B	P	AP
1	-1	-1	-1	1
1	-1	-1	1	-1
1	-1	-1	-1	1
1	-1	-1	1	-1
1	1	-1	1	1
1	1	-1	-1	-1
1	1	-1	-1	-1
1	1	-1	1	1
1	1	1	-1	-1
1	1	1	1	1
1	1	1	-1	-1
1	1	1	1	1
1	-1	1	-1	1
1	-1	1	1	-1
1	-1	1	-1	1
1	-1	1	1	-1

(b) Model matrix

Figure 2.1: Design and model matrices

In Figure 2.1(a), notice that the level settings for factors $A-C$ are listed only four times. This is done to emphasize that the experiment was performed as a split-plot design, with trials performed with factors $A-C$ fixed. The difference between FF and FFSP designs is more obvious when we compare response models. For an FF design, a linear model is usually assumed as:

$$Y = X\boldsymbol{\beta} + \mathbf{e}\ ,$$

where Y is the column vector of response observations, X is the model matrix, $\boldsymbol{\beta}$ is the vector of mean and factorial effects, and $\mathbf{e}$ is the vector of independent random errors that are normally distributed with mean 0 and variance σ^2. In this linear model, there is only one error term. For FFSP designs, on the other hand, the linear model contains two different

source of errors: between plot error and within plot error. To accommodate the two different errors, the linear model for a FFSP design is written as:

$$Y = X\boldsymbol{\beta} + Z\boldsymbol{\varepsilon} + \mathbf{e}\,, \tag{2.1}$$

where Y, X and $\boldsymbol{\beta}$ are as defined previously. The vector $\mathbf{e}$ is a vector of independent SP errors that are normally distributed with mean 0 and variance σ^2_{SP}. The matrix Z indicates to which WP each run belongs to and $\boldsymbol{\varepsilon}$ is a vector of independent WP errors that are normally distributed with mean 0 and variance σ^2_{WP}. The randomization of the experiment and the resulting additional error term are the main features that differentiate a FFSP design from a FF design.

2.2 Example 2

To illustrate another two-level FFSP design, we consider a resistance spot welding example. The original design was described in Bingham (2002) for FFSP variable selection. In this experiment more welds were required to reduce the risk of part failure if the welding process was not reliable and hence, the increase in production cost. They were interested in the impact of abnormal process conditions on the weld quality. The weld quality was evaluated on the strength of the weld. For practical purposes, the quality characteristic in this experiment was the average diameter of the weld button since it was a good surrogate for weld strength. Five factors (shown below as A–E) were investigated, each at two levels. The "High" level indicates ideal operating conditions while the "Low" level indicates an abnormal process condition.

A. Axial misalignment: Incorrect alignment of axes of the upper and lower electrodes

B. Angular misalignment: Deviation of sheet metal parts from the horizontal axis of electrodes

C. Edge Weld: Outer rim of electrodes are lined up with the edge of sheet metal parts

D. Poor Fit-up: Sheet metal parts are not resting together

E. Cooling Water: High and low

To identify conditions that had significant influence on the weld quality, a twelve-run Plackett-Burman design (Plackett and Burman (1946)) was performed. Since it was too time consuming to change the levels of the axial alignment, the experiment was run as a FFSP design with one WP factor, A, and four SP factors, $P - S$. The level setting of axial alignment were only changed once and then fixed throughout the trials while the level settings of the other four factors would be chosen in random order. The design matrix is shown in Figure 2.2.

A	P	Q	R	S
1	1	1	1	1
	−1	−1	1	1
	−1	1	−1	1
	1	−1	1	−1
	1	1	−1	−1
	−1	−1	−1	−1
−1	1	1	1	−1
	−1	−1	1	−1
	−1	1	−1	−1
	1	−1	1	1
	1	1	−1	1
	−1	−1	−1	−1

Figure 2.2: A design matrix of a FFSP design with 1 WP factor and 4 SP factors.

For the 12-runs Plackett-Burman design, each main effect is partially aliased with all 2FI's not containing that main effect. If the effects are computed as the difference between the average response at the high and low levels of a factor, under the model in (2.1), the effect A has variance $2\sigma^2_{WP} + \frac{1}{3}\sigma^2_{SP}$. Each SP main effect has variance $\frac{1}{3}\sigma^2_{SP}$. The variance of the error for an interaction between WP and SP factor is also $\frac{1}{3}\sigma^2_{SP}$ but the variance of the error for an interaction between any two SP factors is $\frac{2}{9}\sigma^2_{WP} + \frac{1}{3}\sigma^2_{SP}$. So, unlike the

completely randomized experiment, the FFSP design has three separate effect variances for the effects instead of one.

Chapter 3

Models, Priors and Model Discrimination Criteria

The main goal of this chapter is to propose a new method for identifying good two-level FFSP designs of different run sizes. We will first introduce the notation used and outline the FFSP model. Then, we will move on to describe the design criterion based on Hellinger distance and the prior distributions for each component in the model that incorporate our prior knowledge regarding the experiment and the three assumptions A1-A3.

3.1 Notation, Model and Design Criterion

In an experiment that has k_1 WP factors and k_2 SP factors, denote W and S as the number of WP and SP settings for a single replicate of the experimental plan. If the WP's do not have the same number of SP trials for a fixed WP setting, then S can be replaced with a vector describing the number of SP settings within each WP. If there are N runs per replicate of the design, and there are R replicates of the experiment plan, we denote the design as $D(N, k_1, k_2, W, S, R)$. When considering an unreplicated design (i.e. $R = 1$), we will refer the design $D(N, k_1, k_2, W, S, 1)$ as $D(N, k_1, k_2, W, S)$. Using our notation, Examples 2.1 and 2.2 are referred to as $D(16, 3, 4, 4, 4)$ and $D(12, 1, 4, 2, 6)$.

A common way to analyze the response from an experiment is to use the linear model (2.1) that contains main effects and 2FI's (we assume that interactions involving three or more interactions are negligible). In this work, a model, M_i, is a collection of $(m_i - 1)$ active factorial effects. The linear model for M_i contains the mean effect and the collection of $m_i - 1$ factorial effects. It is expressed as:

$$Y = X_i\boldsymbol{\beta}_i + Z\boldsymbol{\varepsilon} + \mathbf{e}\ , \tag{3.1}$$

where Y is the $N \times 1$ vector of observations, X_i is the model matrix, $\boldsymbol{\beta}_i$ is the $m_i \times 1$ vector of the mean and factorial effects, and Z is a $N \times (W * R)$ matrix such that the (i, j)-th entry is 1 if the i-th observation appears in the j-th fixed WP and 0 otherwise. The error term, $\boldsymbol{\varepsilon}$, is the $(W * R) \times 1$ vector, which contains independent WP errors that are normally distributed with mean 0 and variance σ^2_{WP}, while $\boldsymbol{e}$ is the $N \times 1$ vector which contains independent SP errors that are normally distributed with mean 0 and variance σ^2_{SP}. Typically $\sigma^2_{WP} > \sigma^2_{SP}$, but this needs not be true in all cases.

Notice that in equation (3.1), the model matrix and linear model coefficients are indexed to distinguish different collections of active effects. The aim of the experiment is to determine, conditional on the observed data, which is the best model. Therefore, we propose to use a design criterion that best distinguishes between plausible models. Intuitively, the best design for an experiment would be able to distinguish all the models of interest. We measure the ability of a design to distinguish between any two models, M_i and M_j, by the Hellinger distances between the predictive densities of the two models, f_i and f_j respectively, given by

$$H(f_i, f_j) = \int (f_i^{\frac{1}{2}} - f_j^{\frac{1}{2}})^2 dY = 2 - 2\int (f_i f_j)^{\frac{1}{2}} dY\ .$$

$H(f_i, f_j)$ measures the distance between the predictive densities of the models M_i and M_j based on the data yielded from a design. The larger $H(f_i, f_j)$ is, the better the design is at distinguishing these two models.

In this work, models are viewed as more likely if they obey the assumptions in A1–A3 and these assumptions are built into the design methodology using a Bayesian framework.

Since there are no observations at this point, the predictive densities are prior distributions of Y, conditional on the model M_i. We defer the discussion of the priors to Section 3.2.

Since we are interested in many competing models, the Hellinger distance between predictive densities of all pairs of model have to be computed when evaluating a design. Bingham and Chipman (2007) proposed the HD-criterion for a particular design, D, as the weighted average of the Hellinger distance between predictive densities of all possible pairs, using the product of the prior probabilities of two models, $P(M_i)P(M_j)$, as weights,

$$C(D) = \sum_{i<j} P(M_i)P(M_j)H(f_i, f_j) \ . \tag{3.2}$$

Not all models are equally likely and models with high probabilities will have a larger influence on the choice of design since the Hellinger distance between their predictive densities are weighted more heavily. A good design will result in a larger value of the HD-criterion, so the HD-criterion is a larger-the-better criterion. It is bounded between 0 and 1 (Bingham and Chipman (2007)) and the bound is useful to identify an optimal run size for a design given the number of factors involved in the experiment (see Section 6.2). The innovation in this work is adapting the HD-criterion to FFSP designs.

3.2 Prior Distributions

The predictive densities used to compute the HD-criterion in equation (3.2) are formulated by a Bayesian framework. The priors are specified in two stages. The first stage is the prior on $\left(\boldsymbol{\beta}_i|\sigma^2_{WP}, \sigma^2_{SP}\right)$ conditional on the model M_i, and the second stage is the prior on the model.

We will first discuss the priors on $\left(\boldsymbol{\beta}_i|\sigma^2_{WP}, \sigma^2_{SP}\right)$ given model M_i that consists of (m_i-1) factorial effects. A conventional independent multivariate normal prior is suggested for $\left(\boldsymbol{\beta}_i|\sigma^2_{WP}, \sigma^2_{SP}\right)$, i.e.

$$\pi\left(\boldsymbol{\beta}_i|\sigma^2_{WP}, \sigma^2_{SP}\right) \sim MVN_{m_i}(\mathbf{0}, (\sigma^2_{WP} + \sigma^2_{SP})\Gamma_i) \ ,$$

where Γ_i is a $m_i \times m_i$ square matrix defined as

$$\Gamma_i = \gamma^2 \begin{pmatrix} c & 0 \\ 0 & I_{m_i-1} \end{pmatrix} ,$$

with I_{m_i-1} as an identity matrix of dimension $(m_i - 1)$. The parameter c that appears in Γ_i is chosen to be $1,000,000$ so that the prior on the intercept has mean 0 and large variance. As suggested by Bingham and Chipman (2007), $\gamma = 2$ is a reasonable uninformative value for starting designs and more careful choice of γ is required for follow-up designs, which will not be considered in this work. Hence, the predictive density of a model M_i is obtained from the predictive distribution of Y, as multivariate normal with mean $\mathbf{0}$ and variance $\sigma^2_{WP}\Sigma_{WP,i} + \sigma^2_{SP}\Sigma_{SP,i}$, where

$$\Sigma_{WP,i} = X_i\Gamma_i X_i^T + Z^T Z ,$$

and

$$\Sigma_{SP,i} = X_i\Gamma_i X_i^T + I_N .$$

The variance term for the predictive distribution of Y given the model, M_i, can be rewritten as

$$\sigma^2_{WP}\Sigma_{WP,i} + \sigma^2_{SP}\Sigma_{SP,i} = \sigma^2_{SP}\left(r\Sigma_{WP,i} + \Sigma_{SP,i}\right) ,$$

where $r = \frac{\sigma^2_{WP}}{\sigma^2_{SP}}$. The reason we want to factor out σ^2_{SP} is such that σ^2_{SP} cancels out in the derivation of the HD-criterion and reduces the computational effort in evaluating the criterion. Following the derivation in Appendix A, the final form of the Hellinger distance between predictive densities is:

$$\int H(f_i, f_j)d\pi(r) = \int 2 - \frac{2}{\left|\frac{1}{2}\left[\Sigma_i^{-\frac{1}{2}}\Sigma_j^{\frac{1}{2}} + \Sigma_i^{\frac{1}{2}}\Sigma_j^{-\frac{1}{2}}\right]\right|^{\frac{1}{2}}} d\pi(r) , \tag{3.3}$$

where $\pi(r)$ is the density of an inverse Gamma distribution and

$$\Sigma_i = r\Sigma_{WP,i} + \Sigma_{SP,i} .$$

Unlike Bingham and Chipman (2007) where the single variance cancels out in the derivation of Hellinger distance between predictive densities, the ratio of the variances is required

to compute the Hellinger distance between predictive densities under the FFSP design. Since, the ratio is unknown, we proposed an inverse Gamma prior with shape parameter $\frac{\nu}{2}$ and scale parameter $\frac{\lambda\nu}{2}$. The choice of λ depends on our prior assumption on the mean of the ratio and ν is determined by the assumption of the quantiles of the ratio. A large ν corresponds to a distribution that is tight around λ while a smaller value will lead to unreasonably heavy tails because $Var(r)$ is undefined for $\nu \leq 4$. Both hyperparameters are chosen carefully to incorporate our prior knowledge of the expected mean and quantiles of the ratio. In the absence of expert knowledge, it is reasonable to assume that the ratio is around 1 and choose $\lambda = 1$, since in most cases the variance within WP is smaller than the variance between the WP's. If expert knowledge is unavailable, the default choice for ν is 5 (Chipman (1998)). More discussion of the choice of the two hyperparameters are can found in Chapter 6.

The prior specification for a model (the collection of active effects) is now considered. This is done by taking into account the assumptions A1–A3. Models with relatively few active effects are preferred. Also, lower order effects are considered to be more likely than higher order effects, and the interaction effects are more likely to have active parent effects.

All the main effects are assumed to have the same probability to be active. Denote this probability as p. This probability can be specified by experimenters. If no such knowledge is available, which is our assumption, p can be determined as described below.

Consider the probability a 2FI is active. Let $p_{AB,0}$, $p_{AB,1}$ and $p_{AB,2}$ denote the conditional probabilities of the 2FI, AB, to be active given that 0, 1 and 2 parent main effects are active, respectively. Effect heredity assumes that 2FI's are more likely to be active if the parent main effects are active. So, $p_{AB,0} \leq p_{AB,1} \leq p_{AB,2}$. In general, the conditional

probability of a 2FI to be active can be described as follows:

$$p_{AB,i} = \begin{cases} c_0 p & \text{if } i = 0 \\ c_1 p & \text{if } i = 1 \\ c_2 p & \text{if } i = 2\ , \end{cases} \tag{3.4}$$

where $0 \leq c_0 \leq c_1 \leq c_2 \leq 1$. The choices of c_i's reflect prior belief in effect hierarchy and heredity. If we feel that, for example, that 2FI's must have both parents in a model, then we would choose $c_0 = c_1 = 0$.

Bingham and Chipman (2007) showed that, for an experiment with k main effects, the expected number of effects under equation (3.4) is

$$E\left[\text{ Number of effects }\right] = kp + \binom{k}{2} p\left[c_0 + 2p(c_1 - c_0) + p^2\left(c_0 - 2c_1 + c_2\right)\right]\ . \tag{3.5}$$

The choice of p is such that the expected number of active effects under the prior matches that of the experimenter's prior belief. For a specified number of effects expected to be active, equation (3.5) can be solved for p. In most situations, it is easier for an experimenter to express belief about the number of anticipated effects than a probability that an effect will be active.

The choice of prior for the model is flexible. Bingham and Chipman (2007) suggested $c_0 = 0.01$, $c_1 = 0.5$ and $c_2 = 1$. Their choice allows the presence of 2FI's in the model even though none of those parents are active with low probability depending on the choice of the experimenters. The model space can be as large as $2^{\frac{k^2+k}{2}}$ for a design that has k main effects. An extreme prior suggested by Meyer et al. (1996) suggested to force a 2FI into the model if both its parents main effects are active but ignore it otherwise, ie. $p_{AB,2} = 1$ and $p_{AB,0} = p_{AB,1} = 0$. Their choice of prior shrinks the model space dramatically and is more computationally appealing. However, the small model space would have limited the robustness of the resulting HD-optimal design. The next chapter explains our approach to expand the model space while keeping the computational costs reasonable.

Chapter 4

Finding Optimal FFSP Designs

The design that maximizes the HD-criterion is viewed as the optimal FFSP design under the proposed framework. The process of finding such a design in the FF setting is not easy. For FFSP designs, finding optimal designs is even more complicated because of the structure of the design. The split-plot structure, with pre-specified numbers of SP settings, must be maintained in the design search.

We first address the issue of computing the Hellinger distance between predictive densities and then move on to explain the search algorithm for HD-optimal FFSP designs. Here, we only consider unreplicated designs that have the same number of SP settings for each WP setting, i.e. $N = W * S$.

When evaluating the Hellinger distance between predictive densities, we need to consider the unknown ratio of variances. This adds another level of difficulty in computing the exact Hellinger distance between predictive densities because the integration does not have a closed form. The integration is approximated using the Monte Carlo method (Gelfand and Smith (1990); Robert and Casella (2005)). Given a random variable X with density $\pi(x)$, the approximation of the integration is

$$\int g(x)d\pi(x) \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i) \ , \tag{4.1}$$

where $x_1, \ldots, x_n$ are randomly generated from the distribution of X. In our case,

$$\int H(r; f_i, f_j) d\pi(r) \approx \frac{1}{n} \sum_{k=1}^{n} H(r_k; f_i, f_j) \ . \tag{4.2}$$

For this application, 1000 random values are generated from the prior distribution of r. By substituting the average of $\int (f_i f_j)^{\frac{1}{2}} dY$ into equation (3.3), we approximate $\int H(f_i, f_j) d\pi(r)$.

In order to calculate the criterion in equation (3.2) for a design, D, we need to evaluate the Hellinger distance between predictive densities for all pairs of models. In most practical applications, there are simply too many models to evaluate. As an example, for a design with 8 factors, we will need to consider 6.8719×10^{10} models. Instead, we choose to evaluate the criterion over models with high prior probabilities. Raftery et al. (1997) too has applied similar approximation for model averaging where they referred to their algorithm as "Occam's Window". To obtain models with higher probability, a large number of models are sampled from the prior distribution and the exact probabilities are evaluated. Factorization of the prior facilitates the simulation of the models, where main effects are first drawn from independent Bernoulli distribution, followed by 2FI's that depend on the main effects. Models that have high prior probabilities are retained for the evaluation of the HD-criterion.

To find optimal designs, an exchange algorithm (Fedorov (1989); Wynn (1972)) is proposed. Since the number of possible designs is too large to search, an evolutionary exchange algorithm similar to that of Boettcher et al. (2000) works well in practice. There are six major steps in our search algorithm.

Step 1 An N-run design, D_0, is generated by randomly generating W WP settings before adding S SP settings to each WP.

We will first begin by generating a random $D(N, k_1, k_2, W, S)$, D_0, as a starting point for our search. The design is generated in two stages. First, W WP settings are selected. Denote them as $WP_1, \ldots, WP_W$. Then, for each WP setting, S SP settings are selected.

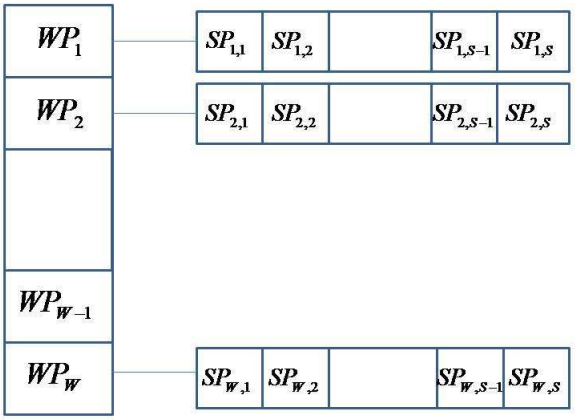


Figure 4.1: Generate design

Denote them as $SP_{i,1}, \ldots, SP_{i,S}$ for the WP setting WP_i. The combination of WP_i and $SP_{i,j}$ is referred to as run (i, j) in the experiment or row (i, j) in the design matrix. Figure 4.1 is a graphical representation of the combination of the W WP settings, and their respective S SP settings. All the settings are generated independently and consist of 1 and -1 to denote the two different levels of the factors.

Step 2 Evaluate the HD-criterion of the design.

The HD-criterion is computed for D_0. Denote this as $C(D_0)$. Note that the same set of models will be used for the implementation of the algorithm.

Step 3 A row is removed from the design matrix and the HD-criterion is evaluated. Denote the criterion as $C(D^-)$. This is repeated for each row in the design matrix. The row, $B-$ that causes the least decrease between $C(D_0)$ and its respective $C(D^-)$ is identified.

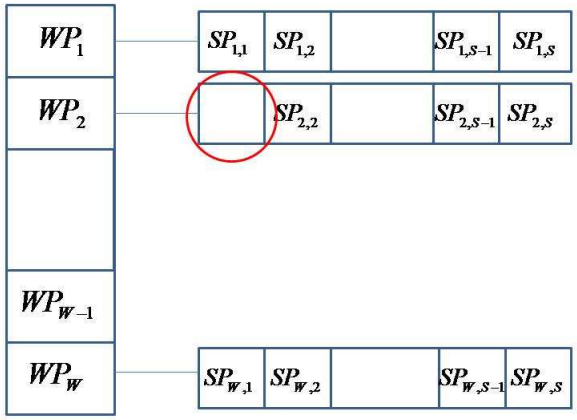


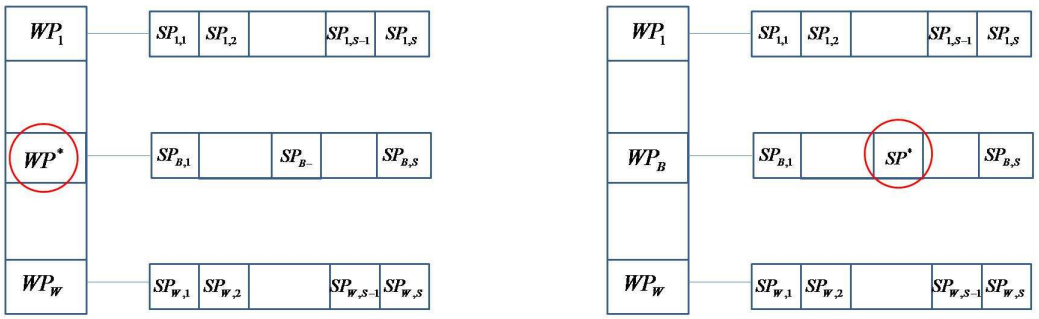
Figure 4.2: Remove one run from the design

The motivation for this step is simple. Each run in a design provides certain amount of information to distinguish between competing models. We would like the runs in an optimal design to be able to provide as much information as possible. Hence, we identify the least informative run in this step and replace it with more informative run in Step 4. To accomplish this, a row is removed from the design matrix and its HD-criterion is calculated, $C(D^-)$. As an example, Figure 4.2 is the graphical representation of row $(2, 1)$ (a combination of WP_2 and $SP_{2,1}$) was removed from the design and its HD-criterion was calculated. This process is repeated for all rows in the design. For each row, we compute $|C(D_0) - C(D^-)|$. The row, $B-$ that causes the least change in the absolute difference between the HD-criteria is identified as the least informative run.

Step 4 Adding a new row to D^-.

Now, we want to replace the least informative run with a more informative one. There are two ways to replace the run: replace the WP settings of all runs that have the same WP setting as $B-$ or replace the SP setting of $B-$. The decision is made by picking the

option that increases the HD-criterion the most. To identify a suitable run, we begin by generating a new WP setting and SP setting independently. Denote them as WP^* and SP^*, respectively. All runs with the same WP settings as $B-$ are replaced with WP^* as shown in Figure 4.3(a). The HD-criterion for the new design is calculated. Then, we proceed to compute the HD-criterion from the new design obtained by only replacing the SP of $B-$ with SP^* as shown in Figure 4.3(b). We continue to generate new WP and SP settings until the new settings cause an increase in the HD-criterion.



(a) Replace WP

(b) Replace SP

Figure 4.3: Replace an informative run

Step 5 Between the two new designs in the previous step, the design that causes the largest increase in the HD-criterion is accepted as a better design.

Lastly, we compare the difference between the two new HD-criteria from Step 4 and $C(D_0)$. The new design that causes the largest increase replaces D_0 in the search and is considered as the optimal design. It will be the new comparison point to decide for the removal and replacement of a run.

Step 6 Steps $2-5$ are repeated until no improvement can be made.

The process of evaluating C(D) for design D, removing one row at a time from the design matrix and replacing the particular row that causes the least decrease in C(D) are repeated alternately until no further changes are made, ie. the new WP and SP settings that maximize HD-criterion is exactly the same as the WP and SP settings of $B-$.

As there is no guarantee that this approach converges to the HD-optimal design, the algorithm is re-started several times, each with a different starting design. In order to maximize computation efficiency, a small set of models are used for preliminary searches of promising designs. Then, each of the promising designs are evaluated using a larger set of models for re-evaluation.

Chapter 5

Examples

There are two examples that will be presented in this chapter to illustrate the methodology. The first example has 6 factors in 16 runs, while the second example considers 11 factors in 12 runs. We will compare the regular design to a MA design while the non-regular design is compared to a D-optimal design.

5.1 A Design of 16 Runs for 6 Factors

Consider an experiment with 6 factors having two WP factors (A and B) and 4 SP factors (P-S) in 16 runs. This is a $D(16, 2, 4, 4, 4)$ FFSP design, where there are 4 WP's and each WP has 4 SP settings. The MA design (Bingham and Sitter (1999)) is a resolution IV design with defining contrast subgroup

$$I = ABPQ = ABRS = PQRS\ . \tag{5.1}$$

The design matrix is shown in Figure 5.1(a). To obtain the HD-optimal design, we assume that the WP and SP variances are almost the same and choose $\lambda = 1$ and $\nu = 5$ such that the ratio of the variances has distribution with mean close to 1 and does not have heavy tails (Chipman (1998)). Further assume that the conditional probability for 2FI's to

A	B	P	Q	R	S
1	1	-1	-1	1	1
		-1	-1	-1	-1
		1	1	1	1
		1	1	-1	-1
-1	-1	-1	-1	1	1
		-1	-1	-1	-1
		1	1	1	1
		1	1	-1	-1
-1	1	-1	1	1	-1
		-1	1	-1	1
		1	-1	1	-1
		1	-1	-1	1
1	-1	-1	1	1	-1
		-1	1	-1	1
		1	-1	1	-1
		1	-1	-1	1

(a) MA design

A	B	P	Q	R	S
1	1	-1	1	-1	-1
		1	1	-1	-1
		1	-1	-1	-1
		1	-1	1	1
-1	-1	-1	1	-1	-1
		1	1	-1	-1
		-1	1	-1	1
		1	1	1	1
-1	1	-1	1	-1	-1
		-1	1	1	-1
		-1	-1	-1	-1
		1	1	1	1
1	-1	-1	1	-1	-1
		1	1	-1	-1
		-1	1	-1	1
		-1	-1	-1	-1

(b) D_5, HD-optimal design

Figure 5.1: FFSP design matrices

be active given that 0, 1 and 2 parent main effects are active is chosen as

$$p_{AB,i} = \begin{cases} p & \text{if } i = 0 \\ 0.5p & \text{if } i = 1 \\ 0.01p & \text{if } i = 2\ . \end{cases} \tag{5.2}$$

The expected number of effect under equations (5.2) and (3.5) is

$$E\left[\text{ Number of effects }\right] = 6p + 30p\left(0.005 + 0.49p + 0.005p^2\right) \tag{5.3}$$

If we expect the number of active effects is 6, then solving for p in equation (5.3) gives us $p = 0.4623$. In the preliminary search, 40 models were used to evaluate the HD-criterion. The resulting designs from 50 random restarts were re-evaluated using 400 models. The resulting best design is D_5, shown in Figure 5.1(b), is different from the MA design. In fact, the HD-optimal design is completely different from the designs that we have seen so

far in the literature of two-level FFSP designs. The difference between the two designs is due to the difference in emphasis of each criterion. The MA criterion attempts to rank designs based on high estimability of all main effects and then followed by higher order interactions. The main effects are not aliased with one another and are also free of aliasing with all the 2FI's. On the other hand, HD-criterion aims to keep aliasing of sets of effects to a minimum. As discussed by Hamada and Wu (1992), non-regular designs such as these can carry more information than regular designs. The reason is that partial aliasing of the non-regular designs can have advantages over regular designs because significant interaction effects could be identified.

Consider the MA designs in (5.1), and the following three models,

$$Y = A + B + AB \ , \tag{5.4}$$

$$Y = A + B + PQ \ , \tag{5.5}$$

and

$$Y = A + B + P + Q + R + S \ . \tag{5.6}$$

Notice that models (5.4) and (5.5) are not distinguishable since $AB = PQ$. Model (5.5), which has a 2FI with no active parent, has prior mass 5.2984×10^{-6}. It is viewed as less likely than model (5.4) that has prior mass 9.8082×10^{-4}. Model (5.6) is viewed as least likely among the three models with prior mass 8.8643×10^{-7}. We take the view that model (5.6) (with all main effects) is unlikely in a screening experiment.

The Hellinger distance between predictive densities is bounded between 0 and 2. Ideally, every model is perfectly distinguishable and hence the Hellinger distance between their predictive densities would be 2. However, this is not possible with such a run size. A HD-optimal design attempts to distinguish as many models as possible, with emphasis on those with high prior probabilities.

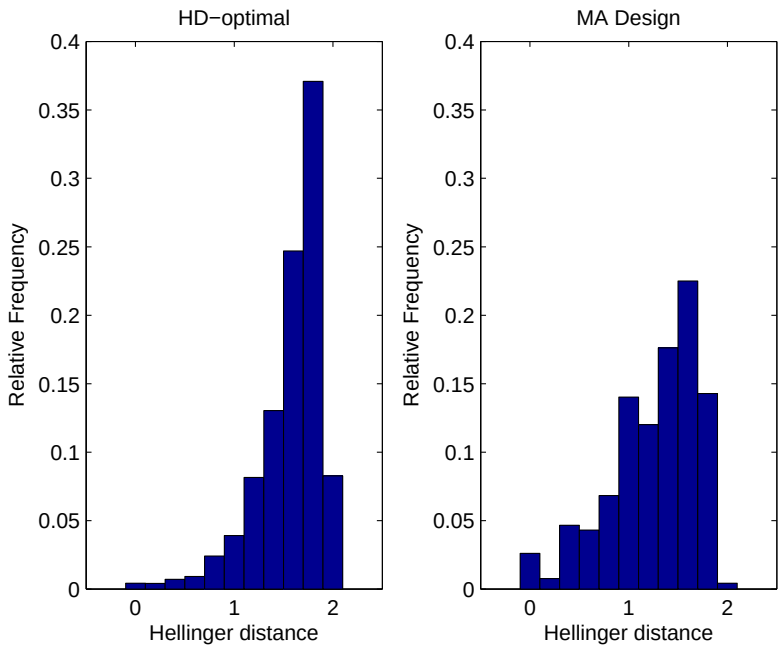


Figure 5.2: Hellinger distance between predictive densities of models that have high probabilities under the HD-optimal and MA designs respectively.

To compare designs, we simulate 400 models with high prior mass as described in Chapter 4. The Hellinger distance between predictive densities for each individual pair of those models is evaluated as in equation (3.3) for the HD-optimal and MA design respectively. The relative frequency histogram of the evaluated Hellinger distance between predictive densities of all pairs of models under the HD-optimal design is shown in histogram on the left of Figure 5.2. The histogram on the right in Figure (5.2) is the relative frequency histogram of the Hellinger distance between predictive densities of each individual pairs of models under the MA design. The relative frequency histograms shows that, there are more pairs of models with large Hellinger distance between predictive densities under the design D_5. Among 79,800 (i.e. $\binom{400}{2}$) pairs of models, D_5 is able to distinguish 84.42% of the individual pairs of models better than the MA criterion design. Among the 40 most likely pairs of models, the HD-optimal design is capable of differentiating the models better than

the MA criterion design 75% of the time.

5.2 12-Run Plackett-Burman Design

Consider a 12-run Plackett-Burman design (Plackett and Burman (1946)) with one WP factor, A, and ten SP factors, P−Y. Kulachi (2007) suggested rearranging the rows of a normal 12-run Plackett-Burman design to obtain a FFSP design that is reasonable if not better when compared to the D-optimal FFSP design generated by SAS JMP®. D-optimality is commonly used as a design criterion. Goos and Vandebroek (2003) also used it for designing FFSP designs. A D-optimal FFSP design is one that maximizes the information matrix, M, defined as

$$M = X^T \Upsilon^{-1} X \ ,$$

where X is the design matrix and Υ is a block diagonal matrix resulting from the Kronecker product of two matrices

$$\Upsilon = I_{W \cdot R} \otimes V \ ,$$

where $I_{W \cdot R}$ is the identity matrix of dimension $W * R$ and V is a $S \times S$ symmetric matrix with $\sigma^2_{WP} + \sigma^2_{SP}$ and σ^2_{WP} as the diagonal and off-diagonal elements respectively.

In order to compare the quality of two designs, Kulachi (2007) used D-efficiency. The D-efficiency is defined as the ratio of the determinants adjusted for the number of parameters estimated, k, based on the main effects models with maximum number of factors that the proposed design can allow. The ratio

$$\left(\frac{|M_1|}{|M_2|} \right)^{\frac{1}{k}} \ ,$$

compares main effects models from two different designs. We use the ratio to compare our proposed design with the D-optimal design shown in Figure (5.3). As mentioned earlier, a design is D-optimal if the determinant of the information matrix is maximized. So, if we let the determinant of the D-optimal design to be the numerator and the determinant of the

A	P	Q	R	S	T	U	V	W	X	Y
-1	1	1	-1	1	1	1	-1	-1	-1	1
	1	-1	1	1	-1	1	1	1	-1	-1
	-1	1	-1	1	1	-1	1	1	1	-1
	-1	-1	1	-1	1	1	-1	1	1	1
	1	1	1	-1	-1	-1	1	-1	1	1
	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	-1	1	1	-1	1	1	1	-1	-1	-1
	-1	-1	-1	1	1	1	1	-1	1	1
	-1	1	1	1	-1	-1	-1	1	-1	1
	1	-1	1	1	1	-1	-1	-1	1	-1
	1	-1	-1	-1	-1	-1	1	1	-1	1
	1	1	-1	-1	1	1	-1	1	1	-1

Figure 5.3: A FFSP design with 1 WP factors and up to 10 SP factors.

HD-optimal design to be the denominator, the larger the ratio, the better is the design in terms of D-optimality.

We want to compare the HD-optimal $D(12, 1, 10, 2, 6)$ FFSP design with the D-optimal design in Figure 5.3. If the expected number of active effects is 5, then $p = 0.2156$ under the prior specification (5.2). There were 40 models used to evaluate the HD-criterion in the exchange algorithm. The resulting designs obtained from 50 random restarts of the exchange algorithm is then re-evaluated using 400 models. Figure 5.4 shows the best among the 50 resulting designs from the random restarts.

The resulting HD-optimal design is again different from the D-optimal design. The adjusted ratio of the determinants shows that the D-optimal design is 310.32% better than the HD-optimal design.

Why are they different? Why is the HD-optimal design not D-optimal? The short answer is that they have different goals. The D-optimal design focuses on the model with 11 main effects. Here, we have taken the view that this model is unlikely. Instead a smaller

A	P	Q	R	S	T	U	V	W	X	Y
-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1
	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	-1	-1	1	-1	-1	-1	-1	-1	-1	-1
	-1	-1	-1	-1	-1	-1	-1	1	-1	-1
	-1	-1	1	-1	-1	-1	-1	-1	1	1
	-1	1	-1	1	-1	-1	-1	-1	1	-1
1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	-1	-1	-1	1	-1	-1	-1	1	-1	-1
	-1	-1	-1	-1	-1	1	-1	1	-1	-1
	-1	1	-1	1	-1	1	1	1	1	-1
	1	-1	1	1	-1	1	1	-1	-1	-1
	1	1	1	1	1	1	1	1	-1	1

Figure 5.4: HD-optimal Design with 1 WP and 10 SP factors.

model with a few interactions is expected.

Similar to Example 5.1, 400 models with high probabilities are simulated. The Hellinger distance between predictive densities of each individual pairs of those models are evaluated as in equation (3.3). Figure 5.5 shows the relative frequency of those Hellinger distance between predictive densities under the HD-optimal and D-optimal design. The relative frequency histogram on the left of Figure 5.5 shows that under the HD-optimal design, there are more pairs of models having large Hellinger distance between predictive densities whereas the opposite is true under the D-optimal design shown in the relative frequency histogram on the right of the same figure. To be precise, among the 79,800 pairs of models, the HD-optimal design is capable of telling apart 88.43% of them better than the D-optimal design.

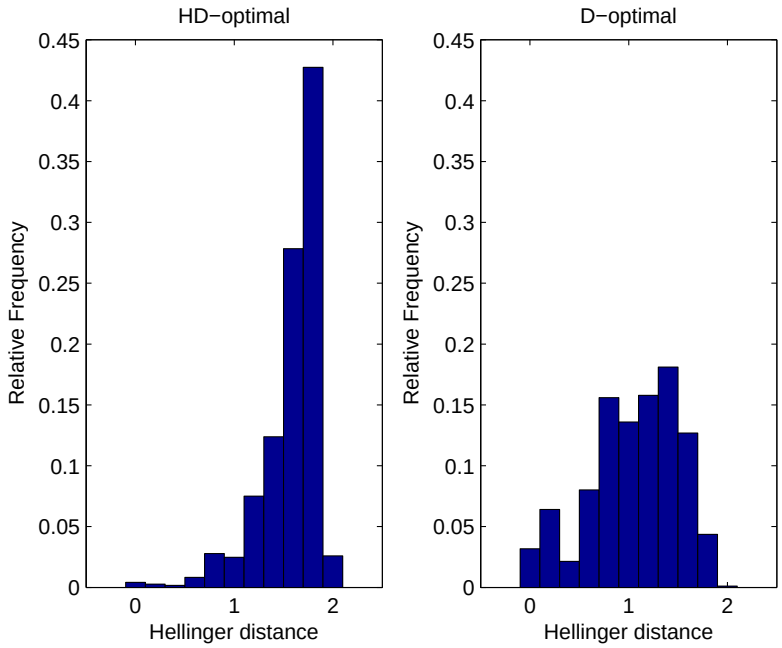


Figure 5.5: Hellinger distance between predictive densities between 79,800 pairs of models under the HD-optimal and D-optimal designs.

Chapter 6

Discussion

In this chapter, we will discuss the flexibility of the methodology with designs from the previous chapter. We will be looking into the selection of hyperparameters and a comparison between balanced and unbalanced designs. The models considered here will also contain only the main effects and the 2FI's.

6.1 Balanced Designs

For a design with an even number of runs, we say that the design is balanced if each factor has an equal number of runs that are set at the same level. The resulting HD-optimal designs shown in Chapter 5 are unbalanced. This is not uncommon in the literature of FFSP designs. Goos and Vandebroek (2003) showed that a three-level D-optimal design for a 42-run experiment suggested by Trinca and Gilmour (2000) is unbalanced. Experimenters are not restricted to use the unbalanced HD-optimal designs. In cases where experimenters are more comfortable with a balanced design, we can limit the search space for HD-optimal designs to balanced designs. As an example, we will shrink the design search space to balanced designs in Example 5.1 to illustrate the flexibility of the criterion. We will use the same parameter and hyperparameters to search for balanced HD-optimal designs. In the preliminary search, we use 40 models and for the final evaluation of the promising designs

from 50 random restarts, 400 models are used. The HD-optimal design obtained, $D_{Balance}$ is shown in Figure 6.1 with $C(D_{Balance})$ 0.7153. The $C(D_5)$ for the unbalanced design is 0.7344.

A	B	P	Q	R	S
1	1	-1	-1	-1	1
		-1	1	-1	1
		1	-1	-1	1
		-1	-1	1	-1
-1	-1	1	-1	1	-1
		-1	-1	-1	-1
		1	1	-1	1
		1	1	-1	-1
-1	1	-1	-1	1	1
		-1	1	1	1
		1	1	1	1
		-1	1	1	-1
1	-1	1	1	1	-1
		1	-1	1	1
		1	-1	-1	-1
		-1	1	-1	-1

Figure 6.1: $D_{Balance}$, a balanced HD-optimal design with 6 factors in 16 runs.

We simulated 400 models independently as described in Chapter 4. The Hellinger distance between predictive densities of each individual pair of models is evaluated under both balanced and unbalanced designs. The histogram on the left in Figure 6.2 shows the relative frequency of the resulting Hellinger distance between predictive densities for all individual pairs of models under the unbalanced design, while the histogram on the right shows the relative frequency of the resulting Hellinger distance between predictive densities for all individual pairs of models under the balanced design. In general, both relative frequency histograms look the same. Close inspection of the figures reveals that the tradeoff is the number of highly distinguishable models under the unbalanced design and the number of

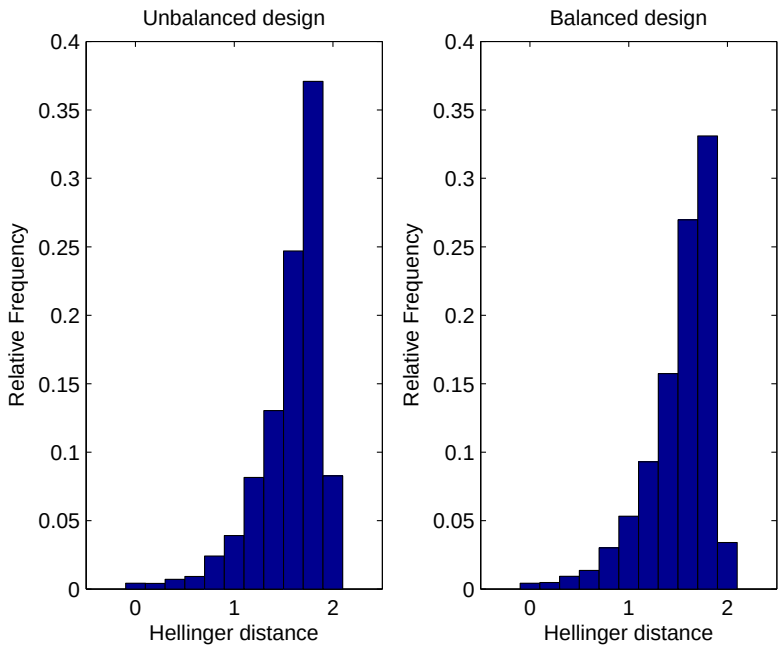


Figure 6.2: Hellinger distance between predictive densities of 79,800 pairs of models.

moderately distinguishable models under the balanced design. If the models we are particularly interested in are more distinguishable under the balanced design, then it is reasonable to choose the balanced design. If there is no specific model that we are interested in, the unbalanced design would be preferred as it is able to better distinguish more of the potential models better. The choice is, again, left to the experimenters to decide the tradeoffs.

6.2 HD-criterion as a Diagnostic Tool

With some assurance of sensible designs, the HD-criterion can also be used to select designs of any desired run size. Bingham and Chipman (2007) showed that by examining the HD-criterion as a function of run size, experimenters could easily choose an appropriate number of runs for their studies. For simplicity, we extend the example 5.1. We want to identify the HD-criterion of the HD-optimal designs for 6 factors, where there are 2 WP

factors and 4 SP factors, with run size ranging from 4 to 32. Four runs are added to the design each time because there are 4 different WP settings. The search for the HD-optimal designs are performed with the same parameter setting from Example 5.1 while changing the choice of ν between 2,5 and 10. Figure 6.3 shows the HD-criteria obtained for each run size. The figure provides the experimenters guidance as to whether the number of trials proposed is sufficient for the experiment given the number of factors involved. The choice of ν does not affect the assessment of optimal run size for an experiment. More discussion of the choice of ν is deferred to Section 6.3. The HD-criterion, when the run size is 1, is obviously 0 because we cannot distinguish between models. As the run size increases from 1, the HD-criterion increases rapidly. With the addition of 4 trials every time, more models with high probabilities can be distinguished. However, after $N = 12$, the rate of increase in the HD-criterion decreases. At this point, almost all probable models are distinguishable. The slight decrease in $C(D)$ when $N = 24$ and $N = 32$ is due to numerical variation.

6.3 Choices of Hyper Parameters

Simulation studies are also performed to understand the effect of the hyper prior parameter ν on the HD-criterion by choosing different values for ν. The simulations are an extension of the previous simulation in Example 5.1 by replacing ν with different values while keeping other parameters the same. The resulting HD-optimal designs are shown in Appendix B. In choosing the prior for r, ν and λ are chosen such that the upper tail contains large value while the middle tail is around the anticipated ratio. λ should be chosen as close to the expected ratio for the variance of the WP and SP factors as possible because for $\nu > 2$, the expected ratio is

$$E(r) = \frac{\lambda\nu}{\nu - 2} \, .$$

The parameter ν acts as the degree of freedom where larger values of ν correspond to distributions that are tighter around λ. Although acceptable, selecting a smaller degree of freedom could lead to an unreasonably heavy tail since $Var(r)$ is undefined when $\nu \leq 4$. It is also reasonable to choose ν such that the upper tail of the distribution of the ratio

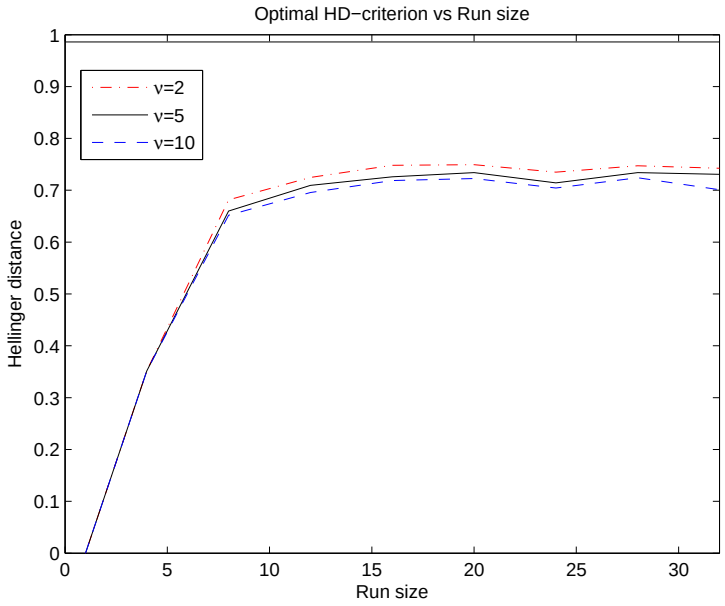


Figure 6.3: C(D) versus the run size of an experiment, for 6 factors with $p = 0.4623$. The upper bound of 0.9864 is also plotted.

is roughly equal to the unconditional ratio. Table 6.1 shows the various quantiles for the inverse Gamma distribution when the parameter λ is fixed at 1.

The choice of ν determines how large the Hellinger distance is between the predictive densities. Smaller ν will result in larger r values being sampled and the Hellinger distance between predictive densities increases proportionally with r. Hence, smaller ν will result in a larger HD-criterion. It is not surprising to see in Figure 6.3 that the $C(D_2)$'s are always consistently larger than $C(D_5)$'s and $C(D_{10})$'s for designs of any run size.

As we take a close look at the design matrices, D_{10} and D_5 are quite similar, but D_2 is different from both of them. The reason for this difference is due to the distribution of r. From Figure 6.4, we see that the distribution of r when $\nu = 2$ has a heavier tail and the expected mean is unknown. It is significantly different from the distributions for r when

Table 6.1: Quantiles of inverse Gamma distribution with $\lambda = 1$

ν	Mean	0.5	0.8	0.9	0.95	0.99
1	-	2.1981	15.5800	63.3281	254.3144	6365.8644
2	-	1.4427	4.4814	9.4912	19.4957	99.4992
3	3	1.2680	2.9846	5.1337	8.5265	26.1252
4	2	1.1916	2.4260	3.7607	5.6281	13.4631
5	1.6667	1.1490	2.1344	3.1050	4.3650	9.0204
6	1.5	1.1219	1.9543	2.7222	3.6689	6.8800
7	1.4	1.1031	1.8313	2.4708	3.2298	5.6495
8	1.3333	1.0893	1.7416	2.2926	2.9276	4.8588
9	1.2857	1.0788	1.6728	2.1593	2.7067	4.3106
10	1.25	1.0705	1.6184	2.0554	2.5379	3.9090

$\nu = 5$ and $\nu = 10$ that is tighter to their respective means.

Although the designs are quite different, all three designs are able to distinguish the models at the similar level of accuracy. This is more obvious when $\nu = 5$ and $\nu = 10$. The designs are able to distinguish the models equivalently well relative to their ability to tell apart other models. To show this, we compare the Hellinger distance between predictive densities of individual pairs of models under all three designs. Among 79,800 pairs of models, the Hellinger distance between predictive densities of the 798 (1%) of the most likely pairs of models are evaluated as in equation (3.3) individually, with $\nu = 2$, $\nu = 5$ and $\nu = 10$ respectively. The resulting Hellinger distance between predictive densities when $\nu = 2$ and $\nu = 10$ are compared to the Hellinger distance between predictive densities of the same pair of models when $\nu = 5$. Figure 6.5(b) shows the resulting Hellinger distance between predictive densities under D_5 plotted against the resulting Hellinger distance between predictive densities under D_{10}. The points follow the straight line $y = x$ closely. This shows that both designs are able to distinguish the same pair of models equivalently well. Figure 6.5(a) shows the resulting Hellinger distance between predictive densities under D_5 plotted against the resulting Hellinger distance between predictive densities under D_2. The points do not follow the straight line $y = x$ but the Hellinger distance between predictive densities

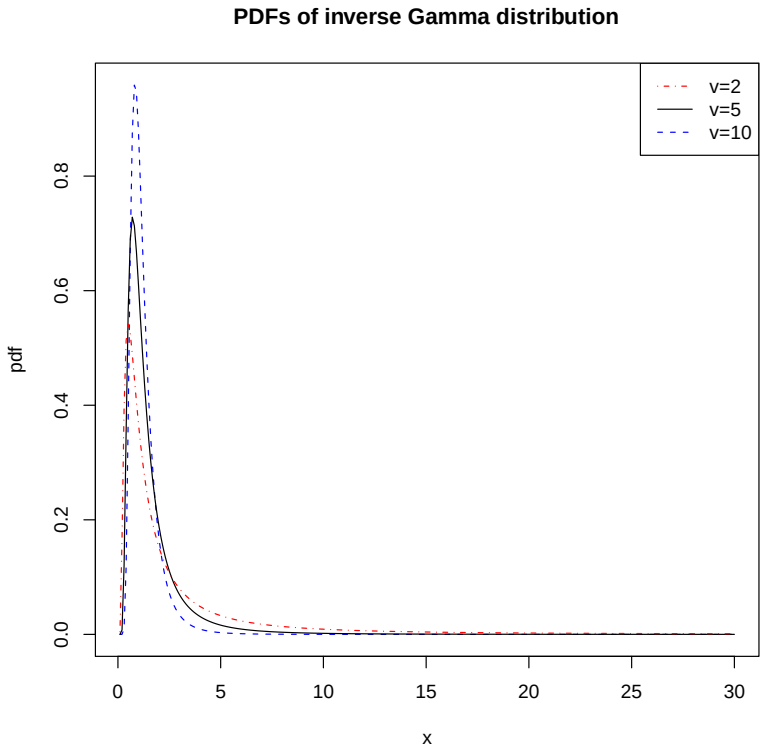
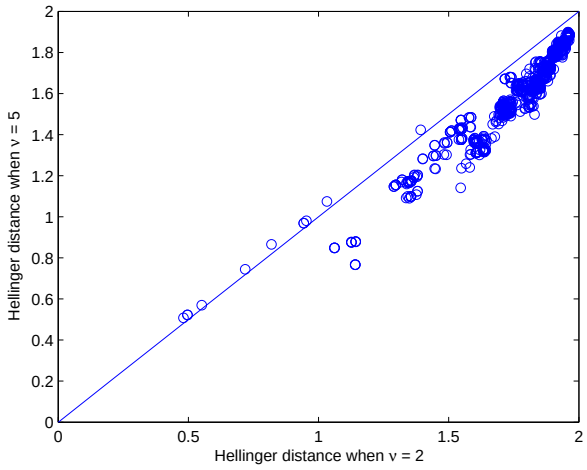


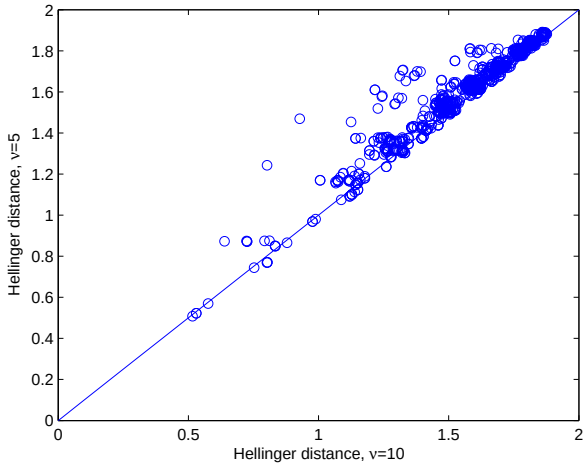
Figure 6.4: Inverse Gamma distribution

of those models under D_5 are positively correlated to the Hellinger distance between predictive densities of the same models under D_2. The Hellinger distances between predictive densities of the probable models are higher for both HD-optimal designs (when $\nu = 2$ and $\nu = 10$) than MA design (83.88% and 85.57% of the time, respectively).

Our methodology can easily be adapted to FF block designs as well. These designs can be viewed as special cases of FFSP designs. Under the FF block designs, interaction effects that define the blocks can be viewed as WP factors while the experiment factors can be viewed as SP factors. Hence, the framework to obtaining HD-optimal FF block designs is similar to obtaining optimal HD-optimal FFSP designs.



(a) Comparison between D_5 and D_2



(b) Comparison between D_5 and D_{10}

Figure 6.5: Comparison of Hellinger distance between predictive densities of the most likely 798 pairs of models under different HD-optimal designs

Chapter 7

Conclusion

In this project, new methodology for finding optimal FFSP designs is developed. We discussed how to overcome the computation challenges when evaluating the criterion and searching for optimal designs. The effectiveness and flexibility of the criterion are illustrated through examples. We demonstrated that HD-optimal designs are able to distinguish between competing models better than designs proposed in the literature.

In this work, we consider only linear models with Gaussian errors. Equation (3.2) is sufficiently general to be applied to other classes of models as well. A natural extension is to consider generalized linear models. Although this is straightforward in practice, some issues, such as the need for approximations to posterior and predictive distributions, can arise.

To summarize, our results include:

- a new criterion for ranking FFSP designs.
- a new algorithm for finding HD-optimal FFSP designs.
- a comparison of HD-optimal designs with designs from the literature.

Appendix A

Derivation of Hellinger Distance

Based on the model and priors mentioned in Section 3, Y is a multivariate normal distribution with mean 0 and variance $\sigma^2_{WP}\Sigma_{WP,i} + \sigma^2_{SP}\Sigma_{SP,i}$, where $\Sigma_{WP,i} = X\Gamma_i X^T + Z^T Z$ and $\Sigma_{SP,i} = X\Gamma_i X^T + I$. Then, the Hellinger distance between predictive densities for two models, M_i and M_j with predictive density f_i and f_j respectively is

$$\begin{aligned} H(f_i, f_j) &= \int (f_i^{\frac{1}{2}} - f_j^{\frac{1}{2}})^2 dY \\ &= 2 - 2\int (f_i f_j)^{\frac{1}{2}} dY \end{aligned}$$

Further simplification of the integration of $(f_i f_j)^{\frac{1}{2}}$ over the data to compute the Hellinger distance between predictive densities yields

$$\begin{aligned} & \int (f_i f_j)^{\frac{1}{2}} dY \\ = & \int \frac{exp\left\{-\frac{1}{2}Y^T\left(\frac{1}{2}\left[\left(\sigma^2_{WP}\Sigma_{WP,i} + \sigma^2_{SP}\Sigma_{SP,i}\right)^{-1} + \left(\sigma^2_{WP}\Sigma_{WP,j} + \sigma^2_{SP}\Sigma_{SP,j}\right)^{-1}\right]\right)Y\right\}}{(2\pi)^{\frac{N}{2}}\left|\sigma^2_{WP}\Sigma_{WP,i} + \sigma^2_{SP}\Sigma_{SP,i}\right|^{\frac{1}{4}}\left|\sigma^2_{WP}\Sigma_{WP,j} + \sigma^2_{SP}\Sigma_{SP,j}\right|^{\frac{1}{4}}} dY \end{aligned}$$

$$
\begin{aligned}
&= \frac{\left|\frac{1}{2}\left[\left(\sigma^2_{WP}\Sigma_{WP,i}+\sigma^2_{SP}\Sigma_{SP,i}\right)^{-1}+\left(\sigma^2_{WP}\Sigma_{WP,j}+\sigma^2_{SP}\Sigma_{SP,j}\right)^{-1}\right]\right|^{\frac{1}{2}}}{\left|\sigma^2_{WP}\Sigma_{WP,i}+\sigma^2_{SP}\Sigma_{SP,i}\right|^{\frac{1}{4}}\left|\sigma^2_{WP}\Sigma_{WP,j}+\sigma^2_{SP}\Sigma_{SP,j}\right|^{\frac{1}{4}}} \\
&\quad \times \int \frac{exp\left\{-\frac{1}{2}Y^T\left(\frac{1}{2}\left[\left(\sigma^2_{WP}\Sigma_{WP,i}+\sigma^2_{SP}\Sigma_{SP,i}\right)^{-1}+\left(\sigma^2_{WP}\Sigma_{WP,j}+\sigma^2_{SP}\Sigma_{SP,j}\right)^{-1}\right]\right)Y\right\}}{(2\pi)^{\frac{N}{2}}\left|\left(\sigma^2_{WP}\Sigma_{WP,i}+\sigma^2_{SP}\Sigma_{SP,i}\right)^{-1}+\left(\sigma^2_{WP}\Sigma_{WP,j}+\sigma^2_{SP}\Sigma_{SP,j}\right)^{-1}\right|^{\frac{1}{2}}}dY \\
&= \frac{\left|\frac{1}{2}\left[\left(\sigma^2_{WP}\Sigma_{WP,i}+\sigma^2_{SP}\Sigma_{SP,i}\right)^{-1}+\left(\sigma^2_{WP}\Sigma_{WP,j}+\sigma^2_{SP}\Sigma_{SP,j}\right)^{-1}\right]\right|^{\frac{1}{2}}}{\left|\sigma^2_{WP}\Sigma_{WP,i}+\sigma^2_{SP}\Sigma_{SP,i}\right|^{\frac{1}{4}}\left|\sigma^2_{WP}\Sigma_{WP,j}+\sigma^2_{SP}\Sigma_{SP,j}\right|^{\frac{1}{4}}} \\
&= \left(\left|\sigma^2_{SP}\Sigma_i\right|^{\frac{1}{4}}\left|\frac{1}{2}\left[\left(\sigma^2_{SP}\Sigma_i\right)^{-1}+\left(\sigma^2_{SP}\Sigma_j\right)^{-1}\right]\right|^{-\frac{1}{2}}\times\left|\sigma^2_{SP}\Sigma_j\right|^{\frac{1}{4}}\right)^{-1} \\
&= \left|\frac{1}{2}\left[\left(\sigma^2_{SP}\Sigma_i\right)^{-\frac{1}{2}}\left(\sigma^2_{SP}\Sigma_j\right)^{\frac{1}{2}}+\left(\sigma^2_{SP}\Sigma_i\right)^{\frac{1}{2}}\left(\sigma^2_{SP}\Sigma_j\right)^{-\frac{1}{2}}\right]\right|^{-\frac{1}{2}} \\
&= \left|\frac{1}{2}\left[\Sigma_i^{-\frac{1}{2}}\Sigma_j^{\frac{1}{2}}+\Sigma_i^{\frac{1}{2}}\Sigma_j^{-\frac{1}{2}}\right]\right|^{-\frac{1}{2}}
\end{aligned}
$$

where $\Sigma_i = r\Sigma_{WP,i} + \Sigma_{SP,i}$ and $r = \frac{\sigma^2_{WP}}{\sigma^2_{SP}}$.

Finally, we need to integrate over r to take into account all possible values of r. The inverse Gamma distribution is chosen as the prior distribution of r but any other proper prior distribution is feasible. If an improper prior is selected, then the Hellinger distance between predictive densities may not be bounded above by 2. Let $\pi(r)$ be the density of r, then substitute the expression in the last step for $\int (f_i f_j)^{\frac{1}{2}} dY$ in the Hellinger distance between predictive densities of the models and integrate over r, we obtain the equation (3.3).

Appendix B

HD-Optimal Designs

A	B	P	Q	R	S
1	-1	-1	1	-1	-1
		1	1	-1	-1
		-1	1	-1	1
		-1	-1	-1	-1
-1	-1	-1	1	-1	-1
		1	1	-1	-1
		-1	1	-1	1
		1	1	1	1
-1	1	-1	1	-1	-1
		-1	1	1	-1
		-1	-1	-1	-1
		1	1	1	1
1	1	1	1	-1	1
		1	-1	-1	1
		1	-1	1	1
		-1	-1	1	-1

Figure B.1: D_2, HD-optimal design when $\nu = 2$.

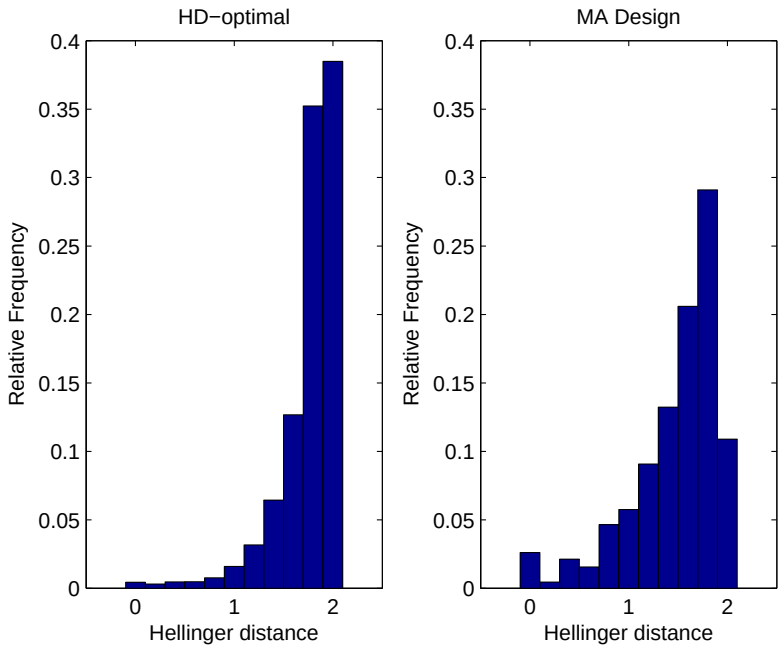


Figure B.2: Relative frequency histograms comparing the Hellinger distance between predictive densities of 79,800 pairs of models when $\nu = 2$ is chosen as the hyper parameter of inverse Gamma distribution. The histogram on the left shows the Hellinger distances between predictive densities under the HD-optimal design while the histogram on the right shows the Hellinger distance between predictive densities under the MA design.

A	B	P	Q	R	S
1	-1	1	-1	-1	1
		1	-1	1	1
		-1	-1	-1	1
		1	-1	-1	-1
-1	-1	1	-1	-1	1
		1	-1	1	1
		-1	-1	-1	1
		-1	-1	1	-1
1	1	1	-1	-1	1
		1	-1	-1	-1
		1	1	-1	-1
		-1	-1	1	-1
-1	1	1	-1	-1	1
		-1	-1	-1	1
		-1	1	-1	1
		1	1	1	-1

Figure B.3: D_{10}, HD-optimal design when $\nu = 10$.

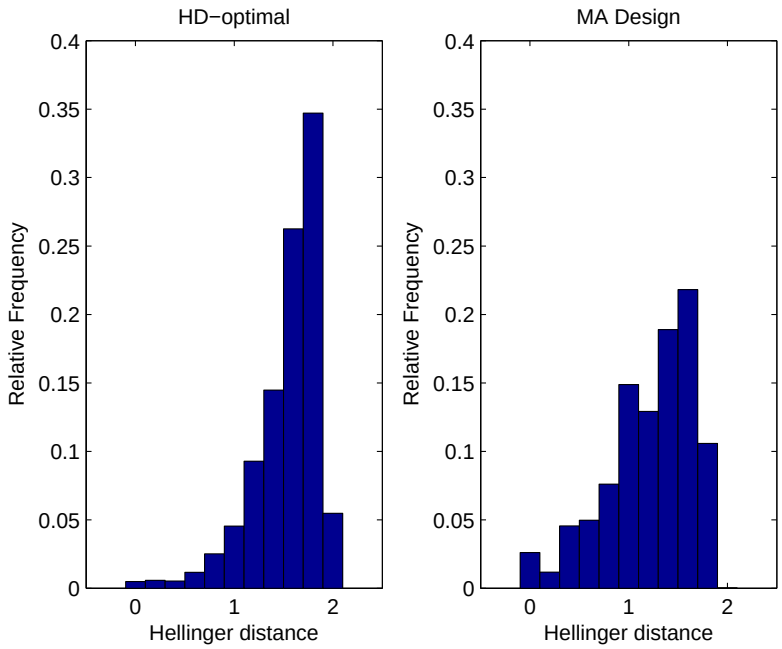


Figure B.4: Relative frequency histograms comparing the Hellinger distance between predictive densities of 79,800 pairs of models when $\nu = 2$ is chosen as the hyper parameter of inverse Gamma distribution. The histogram on the left shows the Hellinger distance between predictive densities under the HD-optimal design while the histogram on the right shows the Hellinger distance between predictive densities under the MA design.

Bibliography

Bingham, D. (2002). Bayesian variable selection for split-plot designs with complex aliasing. *Unpublished* .

Bingham, D. and Chipman, H. (2007). Incorporating prior infomation in optimal design for model selection. *Technometrics* **49**, 155–163.

Bingham, D. and Sitter, R. R. (1999). Minimum aberration two-level fractional factorial split-plot designs. *Technometrics* **41**, 62–70.

Bingham, D. and Sitter, R. R. (2001). Design issues in fractional factorial split-plot experiments. *Journal of Quality Technology* **33**, 2–15.

Bisgaard, S. (2000). The design and analysis of $2^{k-p} \times 2^{q-r}$ split-plot experiments. *Journal of Quality Technology* **32**, 39–56.

Boettcher, S., Percus, A. and Grigni, M. (2000). *Optimizing through Co-evolutionary Avalanches*, volume 1917/2000 of *Lecture Notes in Computer Science*. Springer Berlin, pages 447–456.

Box, G. and Jones, S. (1992). Split-plot designs for robust product experimentation. *Journal of Applied Statistics* **19**, 3–26.

Box, G. E. P. and Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11–18.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in Statistical Analysis*. Addison-Wesley, Co., Massachusetts.

Chen, J., Sun, D. and Wu, C. (1993). A catalogue of two-level and three-level fractional factorial designs with small runs. *International Statistical Review* **61**, 131–145.

Cheng, C., Steinberg, D. and Sun, D. (1999). Minimum aberration and model robustness for two-level fractional factorial designs. *Journal of the Royal Statistical Society* **61**, 85–93.

Cheng, C. and Tsai, P. (2009). Optimal two-level regular fractional factorial block and split-plot designs. *Biometrika* **96**, 83–93.

Cheng, C. S. and Mukerjee, R. (1998). Regular fractional factorial designs with minimum aberration and maximum estimation capacity. *The Annals of Statistics* **26**, 2289–2300.

Chipman, H. (1998). Fast model search for designed experiments. In B. Abraham, editor, *Quality Improvement Through Statistical Methods*. Birkhause Press, pages 207–215.

Chipman, H., Hamada, M. and Wu, C. F. J. (1997). A Bayesian variable selection approach for analyzing designed experiments with complex aliasing. *Technometrics* **39**, 372–381.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics* **1**, 311–341.

Deng, L. and Tang, B. (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular fractional factorial designs. *Statistica Sinica* **9**, 1071–1082.

Fedorov, V. (1989). Optimal design with bounded density: Optimization algorithms of the exchange type. *Journal of Statistical Planning and Inference* **22**, 1–13.

Fries, A. and Hunter, W. (1980). Minimum aberration 2^{k-p} designs. *Technometrics* **22**, 601–608.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistics Association* **85**, 398–409.

Goos, P. and Vandebroek, M. (2003). D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics* **45**, 235–245.

Hamada, M. and Wu, C. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology* **24**, 130–137.

Huang, P., Chen, D. and Voelkel, J. O. (1998). Minimum-aberration two-level split-plot designs. *American Statistical Association* **40**, 314–326.

Kowalski, S. (2002). 24 run split-plot experiments for robust parameter design. *Journal of Quality Technology* **34**, 399–410.

Kulachi, M. (2007). Split-plot experiments with unusual numbers of subplot runs. *Quality Engineering* **19**, 363–371.

Kulachi, M. and Bisgaard, S. (2005). The use of Plackett-Burman designs to construct split-plot designs. *Technometrics* **47**, 495–501.

Kulachi, M., Ramirez, J. and Tobias, R. (2006). Split-plot fractional designs : Is minimum aberration enough? *Journal of Quality Technology* **38**, 56–64.

Li, W. and Wu, J. (1997). Columnwise-pairwise algorithms with applications to the construction of supersaturated designs. *Technometrics* **39**, 171–179.

Meyer, R., Steinberg, D. and Box, G. (1996). Follow-up designs to resolve confounding in multifactor experiments. *Technometrics* **38**, 303–313.

Miller, A. and Sitter, R. (2001). Using the fold-over 12-run Plackett-Burman design to consider interactions. *Technometrics* **43**, 44–55.

Montgomery, D. C. (2005). *Design and Analysis of Experiments*. John Wiley & Sons, INC., New York.

Mukerjee, R. and Fang, K. (2000). Fractional factorial split-plot designs with minimum aberration and maximum estimation capacity. *Statistics Sinica* **12**, 885–904.

Plackett, R. and Burman, J. (1946). The design of optimum multifactorial experiments. *Biometrika* **33**, 305–325.

Raftery, A., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179–191.

Robert, C. and Casella, G. (2005). *Monte Carlo Statistical Methods*. New York: Springer, 2 edition.

Tang, B. and Deng, L. Y. (1999). Minimum G2-aberration for non-regular fractional factorial designs. *The Annals of Statistics* **27**, 1914–1926.

Trinca, L. and Gilmour, S. G. (2000). An algorithm for arranging response surface designs in small blocks. *Computational Statistics and Data Analysis* **33**, 25–43.

Wynn, H. (1972). Results in the theory and consturction of D-optimum experimental designs. *Journal of the Royal Statistical Society* **34**, 133–147.

Xu, H. (2003). Minimum moment aberration for nonregular designs and supersaturated designs. *Statistica Sinica* **13**, 691–708.

Xu, H. and Deng, L. (2005). Moment aberration projection for nonregular fractional factorial designs. *Technometrics* **47**, 121–131.

Yates, F. (1937). The design and analysis of factorial experiments. In *Technical Communication*, volume 35. Commonwealth Bureau of Soil Sciences, London.